AD-A116 432 PENNSYLVANIA STATE UNIV UNIVERSITY PARK DEPT OF STAT--ETC F/G 12/1
RANK-BASED INFERENCE WITHOUT SYMMETRIC ERRORS.(U)
JUN 82 J C AUBUCHON, T P HETTMANSPERGER N00014-80-C-0741
UNCLASSIFIED TR-42 NL

1 OF 1
AD A
116432

END
DATE
FILMED
07-82
DTIC

12

The Pennsylvania State University

Department of Statistics

University Park, Pennsylvania

DEPARTMENT OF STATISTICS
The Pennsylvania State University
University Park, PA 16802, U.S.A.

TECHNICAL REPORTS AND PREPRINTS

Number 42: June 1982

RANK-BASED INFERENCE WITHOUT SYMMETRIC ERRORS

James C. Aubuchon*
The Pennsylvania State University

Thomas P. Hettmansperger**
The Pennsylvania State University

*Currently in the Department of Statistics, Ohio State University

**Research partially supported by ONR contract N00014-80-C-0741

Abstract

Statistical inference based on ranks is reviewed. The role of ~~the~~ parameter $\gamma = \int f^2(x)dx$ and methods for its estimation are discussed. In particular, the use of density estimation methods is shown to provide a consistent estimate ~~$\hat{\gamma}$ of γ~~ γ without the assumption of symmetry of the underlying distribution. The use of a consistent estimate ~~γ~~ in constructing hypothesis tests in the linear model without assuming symmetry is discussed.

1. Introduction

Suppose we have a random sample $X_1, \dots, X_n$ from a distribution with cumulative distribution function (cdf) $F(x-\theta)$. Further, suppose $F(\cdot)$ possesses a symmetric density $f(\cdot)$. Let $\phi^+(u)$, $0 < u < 1$, be a nondecreasing, square-integrable function standardized such that $\int_o^1 [\phi^+(u)]^2 du = 1$. Let $R_1^+, \dots, R_n^+$ be the ranks of the absolute values $|X_1|, \dots, |X_n|$; then

$$S^+ = \sum_{i=1}^n \phi^+\left(\frac{R_i^+}{n+1}\right) \operatorname{sign}(X_i) \tag{1.1}$$

is a rank test statistic for testing H_o: $\theta=0$. The distributional properties of S^+ were studied in great detail by Hajek and Sidak (1967). The test rejects H_o: $\theta=0$ in favor of H_A: $\theta\neq0$ if $|S^+| \geq c$, where c may be determined from the *permutation distribution* of S^+ or a normal approximation. Under H_o: $\theta=0$, $n^{-1/2} S^+$ has an asymptotic standard normal distribution. The Hodges-Lehmann (1963) estimate of θ is an approximate solution of $S^+(\theta)=0$, where $S^+(\theta) = \sum_{i=1}^n \phi^+(R_i^+(\theta)/(n+1))\operatorname{sign}(X_i-\theta)$ and $R_i^+(\theta)$ is the rank of $|X_i-\theta|$, $i=1, \dots, n$. If $P_{H_o}(S^+ \leq -k) = \alpha/2$ then $[\hat{\theta}_L, \hat{\theta}_U]$ is a $(1-\alpha)$ 100% confidence interval for θ, where $S^+(\hat{\theta}_L) = k$ and $S^+(\hat{\theta}_U) = -k$ define the end points.

Define

$$\phi^+(u,f) = -\frac{f'\{F^{-1}[(u+1)/2]\}}{f\ \{F^{-1}[(u+1)/2]\}}\ , \tag{1.2}$$

and

$$\tau^{-1} = \int_o^1 \phi^+(u)\ \phi^+(u,f)du. \tag{1.3}$$

Then the Pitman efficacy of the test based on S^+ is τ^{-1} (Hajek and Sidak 1967, p 220). Hence, the efficiency properties of S^+ are determined by τ^{-1}. Further, the estimate $\hat{\theta}$ has a normal limiting distribution with mean θ and variance τ^2/n (Hodges and Lehmann 1963). Finally,

$$\hat{\tau} = (\hat{\theta}_U - \hat{\theta}_L)/(2k) \tag{1.4}$$

converges in probability to τ (Sen 1966). Hence, if the efficiencies of the point estimate and confidence interval are defined by their asymptotic variance and asymptotic length, then the remarks above show that estimation methods inherit their efficiency properties from S^+, the parent test statistic. We now turn to the linear model.

Let $Y_1, \ldots , Y_n$ be independent observations on a linear model. We suppose Y_i has cdf $F(y-\alpha-x_i'\beta)$, $i=1, \ldots n$, where β is a pxl vector of regression parameters, x_i' is the i^{th} row of the nxp design matrix X, α is an intercept parameter and $F(\cdot)$ has density $f(\cdot)$. We define a score function $\phi(\cdot)$ corresponding to the one-sample score function $\phi^+(\cdot)$ by

$$\begin{array}{rll} \phi(u) = & -\phi^+(1-2u) \quad , & 0 < u \leq 1/2 \\ & \phi^+(2u-1) \quad , & 1/2 < u < 1. \end{array} \tag{1.5}$$

Then $\int_o^1 \phi(u)du = 0$ and $\int_o^1 \phi^2(u)du = 1$. Let

$$D(y-\alpha-X\beta) = \sum_{i=1}^{n} \phi(\frac{R_i(\beta)}{n+1})(Y_i-\alpha-x_i'\beta) \; , \tag{1.6}$$

where $R_1(\beta), \cdots , R_n(\beta)$ are the ranks of $Y_1-\alpha-x_1'\beta, \ldots , Y_n-\alpha-x_n'\beta$. Then Jaeckel (1972) defined a rank estimate of β (D is invariant to α) by a value

$\hat{\beta}$ which minimizes $D(y-\alpha-X\beta)$. The negative gradient of D with respect to β exists for almost all β and is equal to the px1 vector $S(Y-\alpha-X\beta)$with j^{th} component $S_j(Y-\alpha-X\beta) = \sum_{i=1}^{n} x_{ij}\phi(R_i(\beta)/(n+1))$. An asymptotically equivalent rank estimate $\hat{\beta}$ may be defined by $S(Y-\alpha-X\beta) \doteq 0$. Jureckova (1971) and Jaeckel (1972) showed that $\hat{\beta}$ has a multivariate normal limiting distribution with mean β and covariance matrix $\tau^2(X_c{}'X_c)^{-1}$, where X_c is the centered design matrix, assumed to have full rank.

The parameter τ can be defined in terms of the score function $\phi(\cdot)$ as follows: let $\phi(u,f)$ be derived from $\phi^+(u,f)$ through 1.5; then

$$\int_0^1 \phi(u)\ \phi(u,f)du = \int_0^1 \phi^+(u)\ \phi^+(u,f)du = \tau^{-1}.$$

If we write $X\beta = X_1\beta_1 + X_2\beta_2$, where β_1 is (p-q)x1 and β_2 is qx1, then we can consider rank-based tests of H_o: $\beta_2=0$, β_1 unspecified. Let $\hat{\beta}_1$ be the rank estimate of β_1 when $\beta_2=0$, (reduced-model estimate), and let $\hat{\beta}$ denote the full-model estimate of β. Partition S into S_1 and S_2; then the qx1 vector of aligned rank statistics $S_2(Y-\alpha-X_1\hat{\beta}_1)$, under H_o, has a multivariate normal limiting distribution with mean 0 and a covariance matrix which can be written using a similar partitioning of the centered design X_c:

$$\Lambda = n^{-1}[X_{2c}{}'X_{2c} - X_{2c}{}'X_{1c}(X_{1c}{}'X_{1c})^{-1}X_{1c}{}'X_{2c}] \ . \qquad (1.7)$$

See Adichie (1978) and Sen and Puri (1977).

Under H_o the following three random variables all have asymptotic chi-square distributions with q degrees of freedom:

$$S_2{}'(Y-\alpha-X_1\hat{\beta}_1)\Lambda^{-1}S_2(Y-\alpha-X_1\hat{\beta}_1) \ , \qquad (1.8)$$

$$\frac{(H\hat{\beta})'[H(X_c'X_c)^{-1}H']^{-1}(H\hat{\beta})}{\tau^2} , \qquad (1.9)$$

where H = (O,I) such that $H\beta=\beta_2$, and

$$\frac{D(Y-\alpha-X_1\hat{\beta}_1) - D(Y-\alpha-X\hat{\beta})}{\tau/2} . \qquad (1.10)$$

See McKean and Hettmansperger (1976) for a discussion of 1.10. Tests based on 1.8, 1.9 and 1.10 have the same Pitman efficacy and so cannot be separated using asymptotic efficiency.

For a moment suppose τ is known. Then the asymptotic distribution theory for 1.8-1.10 does not require symmetry of $f(\cdot)$. However, if τ must be estimated or if we wish to make inferences on the intercept α, then one-sample methods seem to be needed. First form the full-model residuals $r_i = y_i - x_i'\hat{\beta}$, $i=1, \ldots , n$. Now apply S^+ to estimate α and to compute $\hat{\tau} = (\hat{\alpha}_U-\hat{\alpha}_L)/(2k)$. McKean and Hettmansperger (1976) showed that $\hat{\tau}$ is consistent, provided $f(\cdot)$ is symmetric.

Hence, if we suppose $f(\cdot)$ is symmetric, then we can estimate τ consistently for use in constructing the tests based on 1.9 and 1.10 and for estimating the asymptotic standard errors of $\hat{\beta}$. In the remainder of the paper we consider what happens to $\hat{\tau}$ when $f(\cdot)$ is not symmetric and discuss alternative estimates of τ which do not require symmetry.

2. Estimation of $\int f^2(x)dx$

In this section we will consider a random sample $X_1, \ldots, X_n$ from a distribution with cdf F(x). Further, suppose F(•) possesses a density f(•) which may not be symmetric. We will restrict attention to Wilcoxon scores defined by $\phi^+(u) = 3^{1/2}u$. Then 1.3 becomes $\tau^{-1} = 12^{1/2}\int_{-\infty}^{\infty} f^2(x)dx$, and we will concentrate on the consistent estimation of

$$\gamma = \int_{-\infty}^{\infty} f^2(x)dx. \qquad (2.1)$$

When f(•) is symmetric, the consistency of $\hat{\tau}$ in 1.4 is usually derived from an asymptotic linearity result on S^+ due to van Eeden (1972). The result can be easily modified for f(•) with arbitrary shape:

$$n^{-1/2}S^+(n^{-1/2}b) - n^{-1/2}S^+(n^{-1/2}a) = -(b-a)(12)^{1/2}\int_{-\infty}^{\infty} f(-x)f(x)dx + o_p(1), \qquad (2.2)$$

where $o_p(1)$ tends to zero in probability uniformly for all a,b such that $|b-a| \leq K$, for any positive constant K. By $S^+(\theta)$ we mean S^+ computed on $X_i-\theta$, i=1, ..., n.

Since $n^{1/2}\hat{\theta}_L$ and $n^{1/2}\hat{\theta}_U$ are bounded in probability, where $S^+(\hat{\theta}_L) = k$ and $S^+(\hat{\theta}_U) = -k$, we have from 2.2

$$\hat{\gamma} = -\frac{S^+(\hat{\theta}_U) - S^+(\hat{\theta}_L)}{12^{1/2}n(\hat{\theta}_U - \hat{\theta}_L)} \qquad (2.3)$$

$$= \frac{2k}{12^{1/2}n(\hat{\theta}_U-\hat{\theta}_L)}$$

$$= \int_{-\infty}^{\infty} f(-x)\ f(x)dx + o_p(1)\ .$$

Note $k \doteq Z_{\alpha/2}n^{1/2}$ from the normal approximation where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution. Note that k is chosen under the assumption of symmetric $f(\cdot)$ and in general will not be the $\alpha/2$ critical value of S^+. This does not effect the convergence in 2.3. From 2.3, since $\hat{\gamma} \xrightarrow{P} \int f(-x)f(x)dx$, it is evident that the usual estimate $\hat{\tau}$, 1.4, is no longer consistent. The amount of large sample bias depends on the size of $\int f^2(x)dx - \int f(-x)f(x)dx$.

We next consider Jaeckel's (1971) model of asymmetric contamination. Suppose for a given sample size n we are sampling from

$$G_n(x) = (1-cn^{-1/2})F(x) + cn^{-1/2}H(x) \tag{2.4}$$

where $F(\cdot)$ is the cdf of a symmetric distribution and $H(\cdot)$ is the cdf of a distribution with arbitrary shape. Jaeckel points out: "The amount of asymmetric contamination is large enough to affect the performance of the estimator, but is too small to be measured accurately at the given sample size." We will present a heuristic derivation to suggest that $n^{1/2}(\hat{\gamma} - \int f^2(x)dx)$ has an asymptotic $n(b,\sigma^2)$ distribution where b is the asymptotic bias and $b^2+\sigma^2$ is the asymptotic mean square error. Thus, both the bias and variability of the estimator approach zero at the same rate.

Using the heuristic argument of Huber (1969) to construct the projection of the estimator, we have

$$n^{1/2}(\hat{\gamma} - \int f^2(x)dx) = 2\ n^{-1/2} \sum_{i=1}^{n} [f(X_i) - \int f^2(x)dx] + o_p(1) \tag{2.5}$$

Under the model of symmetry, with no contamination, $n^{1/2}(\hat{\gamma} - \int f^2(x)dx)$ is asymptotically $n(0,\ 4\{\int f^3(x)dx - [\int f^2(x)dx]^2\})$. For a rigorous discussion see Antille (1972, 1974).

To determine the limiting distribution of 2.5 under the contamination model 2.4 we use the contiguity results of Hajek and Sidak (1967, Chapter 6). We will assume sufficient smoothness of the densities f(•) and h(•) to carry out the required expansions. Then the log-likelihood can be written as

$$\log L = cn^{-1/2} \sum_{i=1}^{n} \left(\frac{h(X_i)}{f(X_i)} - 1\right) + o_p(1). \tag{2.6}$$

Now $(n^{1/2}(\hat{\gamma} - \int f^2(x)dx), \log L)$ is asymptotically bivariate normal. We need the covariance:

$$b = 2cE\{[f(X_i) - \int f^2(x)dx][\frac{h(X_i)}{f(X_i)} - 1]\} \tag{2.7}$$

$$= 2c[\int f(x)h(x)dx - \int f^2(x)dx] \ .$$

The above results show that the densities $\Pi g_n(x_i)$ and $\Pi f(x_i)$ are contiguous, and the limiting distribution of $n^{1/2}(\hat{\gamma} - \int f^2(x)dx)$, under the contamination model 2.4, is $n(2c[\int f(x)h(x) - \int f^2(x)dx], 4[\int f^3(x)dx - (\int f^2(x)dx)^2])$. The expression 2.7 then represents the asymptotic bias, which can be far from zero, depending on the choice of h(•).

We now return to 2.3 for a different representation of the estimator $\hat{\gamma}$. Recall the counting form for the Wilcoxon signed-rank statistic $S^+(\theta)$:

$$S^+(\theta) = \frac{3^{1/2}}{n+1}\left\{\sum_{i\leq j}\sum I(\frac{X_i + X_j}{2} > \theta) - \sum_{i\leq j}\sum I(\frac{X_i + X_j}{2} < \theta),\right\} \tag{2.8}$$

$$= \frac{2(3^{1/2})}{n+1}\left\{\sum_{i\leq j}\sum I(\frac{X_i + X_j}{2} > \theta) - \frac{n(n+1)}{2}\right\}$$

Let $T(\theta) = [n(n+1)]^{-1} \sum_{i\leq j}\sum I(X_i+X_j > 2\theta)$; then from 2.3

$$\hat{\gamma} = -\frac{T(\hat{\theta}_U) - T(\hat{\theta}_L)}{\hat{\theta}_U - \hat{\theta}_L} \quad . \tag{2.9}$$

$$= \frac{T(\hat{\theta}_L) - T(\hat{\theta}_U)}{h_n/2} \quad ,$$

where $h_n = 2(\hat{\theta}_U - \hat{\theta}_L) = 2(\hat{\theta}_U - \hat{\theta})$, $\hat{\theta} = (\hat{\theta}_L + \hat{\theta}_U)/2$. Note that $h_n \xrightarrow{P} 0$ and $\hat{\theta} \xrightarrow{P} \theta$, which we take to be zero without loss of generality. With a bit of algebra, we can write 2.9 as

$$\hat{\gamma} \doteq \frac{1}{n(n+1)h_n} \sum_{i \neq j}\sum I\{|X_i + X_j - 2\hat{\theta}| < \frac{h_n}{2}\} \tag{2.10}$$

$$+ \frac{2}{n(n+1)h_n} \sum_i I\{|X_i - \hat{\theta}| < \frac{h_n}{2}\} \ .$$

The approximate equality is due to using absolute values to represent the counts in 2.8. We next show that $\hat{\gamma}$ can be related to density estimators.

Rosenblatt (1956) proposed the following simple, rectangular window estimator of f(x):

$$f_n(x) = \frac{1}{nh_n} \sum I\{|x - X_i| < \frac{h_n}{2}\} \ . \tag{2.11}$$

For large n, $\hat{\gamma}$ is essentially the same as

$$\gamma^* = \frac{1}{n^2 h_n} \sum_{i \neq j}\sum I\{|X_i + X_j| < \frac{h_n}{2}\} \tag{2.12}$$

$$= \int_{-\infty}^{\infty} f_n(-x)\, dF_n(x) \ ,$$

where $F_n(x)$ is the empirical cdf. This representation suggests two points: (1) To estimate $\int f^2(x)dx$ consistently, we should consider pairwise differences rather than pairwise sums or averages, and (2) it may be advantageous to use density estimation directly. In the next section we discuss density estimation of $\int f^2(x)dx$. For estimates based on the differences the reader is referred to Sievers (1982).

3. Density Estimates of $\int f^2(x)dx$

As suggested by 2.12, consistent estimation of γ in 2.1, for $f(\cdot)$ with arbitrary shape, may be achieved by using density estimates. Suppose that $W(\cdot)$ is a given density, symmetric about 0. Let

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} W(\frac{x-X_i}{h_n}) .$$

Then $f_n(x)$ is called a window estimate of $f(x)$, $W(\cdot)$ is called the window, and $h_n \to 0$ is called the window width. These estimates were first discussed by Rosenblatt (1956) and Parzen (1962). See Wegman (1972) and Bean and Tsokas (1980, 1982) for reviews of density estimation. The estimate 2.11 is an example, with $W(\cdot)$ taken as the uniform density on (-1, 1).

Bhattacharyya and Roussas (1969) considered $\int f_n^2(x)dx$, and Cheng and Serfling (1981) considered the more general $\int \phi(x)\psi(\hat{F}(x)f_n^2(x)dx$, where $\hat{F}(x)$ is the integral of $f_n(x)$. On the other hand, Schuster (1974) and Ahmad (1976) studied $\int f_n(x)dF_n(x)$, where $F_n(x)$ is the empirical cdf. Schweder (1975) considered $\int \psi(F_n(x))f_n(x)dF_n(x)$. The more general form of each approach is appropriate for estimation of τ^{-1} in 1.3. Again, we restrict attention to the special case of γ, 2.1 The works cited above contain results on the weak and strong consistency and weak convergence of the various estimates. The following theorem relates the two estimators, $\int f_n^2(x)$ and $\int f_n(x)dF_n(x)$.

Theorem. Let $f_n(\cdot)$ be given by 3.1. Then $\int f_n{}^2(x)dx = \int f_n{}^*(x)dF_n(x)$, where $f_n{}^*(x)$ is defined by 3.1 using $W^*(x) = W*W(x)$, the convolution density.

Proof. We have

$$\int f_n{}^2(x)dx = \int (nh_n)^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} W(\frac{X_i-x}{h_n})W(\frac{X_j-x}{h_n})\ dx.$$

Let $z = (X_j-x)/h_n$; then

$$\int f_n{}^2(x)dx = n^{-2}h_n{}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \int W(z + \frac{(X_i-X_j)}{h_n})W(z)dz$$

$$= n^{-2}h_n{}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} W^*(\frac{X_i-X_j}{h_n}) \ ,$$

where $W^*(z)$ is the convolution of $W(z)$ with $W(-z)$. But since $W(\cdot)$ is symmetric about 0, $W^*(z) = W*W(z)$. Now let $f_n{}^*(x)$ be given by 3.1, with window $W^*(\cdot)$; then $\int f_n{}^2(x)dx = \int f_n{}^*(x)dF_n(x)$, where $F_n(\cdot)$ is the empirical cdf.

Since not every density can be written as the convolution of some density with itself, we cannot, in general, reverse the argument. Further the relationship seems only to hold for the case of Wilcoxon scores.

In the following discussion we present still another motivation for using $\int f_n(x)dF_n(x)$ as a natural estimate of $\int f^2(x)dx$. Note that

$$P(X_1-X_2 < q) = \int_{-\infty}^{\infty} \int_{-\infty}^{y+q} dF(x)dF(y). \qquad (3.2)$$

Define the functional T(F) by

$$T(F) = \frac{d}{dq} \int_{-\infty}^{\infty} \int_{-\infty}^{y+q} dF(x)dF(y) \Bigg|_{q=0} \qquad (3.3)$$

$$= \int_{-\infty}^{\infty} f^2(x)dx \ .$$

This suggests the estimate $T(F_n)$, which can be approximated by

$$\hat{T}(F_n) = q_n^{-1}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{y+q_n} dF_n(x)dF_n(y) - \int_{-\infty}^{\infty}\int_{-\infty}^{y} dF_n(x)dF_n(y)\right] , \qquad (3.4)$$

where $q_n \to 0$ $(q_n > 0)$. This reduces to

$$\hat{T}(F_n) = (2q_n n^2)^{-1}\sum_{i\neq j}\sum I\{|X_i - X_j| \le q_n\} \qquad (3.5)$$

$$\doteq \int f_n(x)dF_n(x) \ ,$$

with $h_n = 2q_n$. The approximation arises because the i=j terms are not present on the left-hand side and we have $\le$ in the indicator.

Let $F_\varepsilon(x) = (1-\varepsilon)F(x) + \varepsilon\delta_z(x)$, where $\delta_z(x)$ is the cdf which puts mass 1 at z. Then Hampel's (1974) influence curve for the functional $T(F)$ is defined as

$$IC(z) = \lim_{\varepsilon\to 0}\frac{T(F_\varepsilon)-T(F)}{\varepsilon} \ . \qquad (3.6)$$

From 3.3 we have

$$T(F_\varepsilon) = \frac{d}{dq}\int_{-\infty}^{\infty}[(1-\varepsilon)F(y+q) + \varepsilon\delta_z(y+q)]d[(1-\varepsilon)F(y) + \varepsilon\delta_z(y)]\Big|_{q=0}$$

$$= (1-\varepsilon)^2\int_{-\infty}^{\infty}f^2(x)dx + 2\varepsilon(1-\varepsilon)f(z).$$

Hence,

$$IC(z) = 2(f(z) - \int f^2(x)dx), \qquad (3.7)$$

and the influence is bounded provided the density $f(\cdot)$ is bounded.

Following the heuristic argument of Huber (1981, p 14) the functional $T(F_n)$, expanded about F, yields

$$n^{1/2}[T(F_n) - \int f^2(x)dx] = 2\ n^{-1/2} \sum_{i=1}^{n} [f(X_i) - \int f^2(x)dx] + R_n. \qquad (3.8)$$

Provided $R_n \xrightarrow{P} 0$, the leading term in 3.8 shows that we can anticipate, from the Central Limit Theorem, that $n^{1/2}[T(F_n) - \int f^2(x)dx]$ is asymptotically $n(0,\ 4\{\int f^3(x)dx - [\int f^2(x)dx]^2\})$. See the references at the beginning of this section for a rigorous development. Hence, we have the same limiting distribution for the density type estimate for the case of arbitrary shape as for $\hat{\gamma}$ in the symmetric case described by 2.5.

Anticipating the estimation problems in the linear model raised in the Introduction, we discuss the estimation of γ in the two-sample location model. This model is the simplest linear model and provides a framework for comparing the confidence interval approach to the density estimation approach.

Suppose that we observe $Y_1, \dots Y_n$, with $Y_1, \dots Y_m$ and $Y_{m+1}-\beta, \dots Y_n-\beta$ all iid with continuous cdf $F(\cdot)$. We suppose, without loss of generality, that $F(0) = 1/2$. This corresponds to a linear model with $\alpha=0$ and X an nxl vector of m zeros and n-m ones, so that $Y = X\beta+e$ as described in the Introduction. Further, with $\phi^+(u) = 3^{1/2}u$, from 1.5 we have $\phi(u) = (12)^{1/2}(u-1/2)$, the two-sample Wilcoxon score-function. The gradient of D in 1.6 suggests the Mann-Whitney-Wilcoxon statistic:

$$S(\beta) = \frac{(12)^{1/2}}{n+1} \sum_{i=m+1}^{n} [R_i(\beta) - (n+1)/2], \qquad (3.9)$$

where $R_1(\beta), \dots, R_n(\beta)$ are the ranks of $Y_1, \dots Y_m, Y_{m+1}-\beta, \dots Y_n-\beta$. (We will treat $Y_1, \dots Y_m$ as the first sample and $Y_{m+1}, \dots Y_n$ as the second).

The random variable $S(\beta)$ can be written in terms of counts as

$$S(\beta) = \frac{(12)^{1/2}}{n+1} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \{I(Y_j - Y_i > \beta) - 1/2\} . \tag{3.10}$$

Suppose $P(S(0) \leq -k) = \alpha/2$. Then $S(\hat{\beta}_L) = k$ and $S(\hat{\beta}_U) = -k$ define $[\hat{\beta}_L, \hat{\beta}_U]$, a $(1-\alpha)$ 100% confidence interval for β. Suppose $m/n \to \lambda$, $0 < \lambda < 1$; then Lehmann (1963) showed that

$$\hat{\gamma}_L = \frac{2k}{[12\ \lambda(1-\lambda)]^{1/2}[m(n-m)]^{1/2}(\hat{\beta}_U - \hat{\beta}_L)} = \frac{2Z_{\alpha/2}}{12\ \lambda(1-\lambda)^{1/2}(n+1)^{1/2}(\hat{\beta}_U - \hat{\beta}_L)} \tag{3.11}$$

$$\xrightarrow{P} \int f^2(x)dx,$$

where $k \doteq Z_{\alpha/2}\ (m(n-m)/(n+1))^{1/2}$ from the normal approximation, similar to 2.3.

Now let $h_n = \hat{\beta}_U - \hat{\beta}_L$ and $\hat{\beta} = (\hat{\beta}_U + \hat{\beta}_L)/2$. Clearly, $h_n \xrightarrow{P} 0$ and $\hat{\beta} \xrightarrow{P} \beta$. Since $n^{1/2}(\hat{\beta} - \beta)$ is bounded in probability, using 3.10 we can rewrite 3.11 as

$$\hat{\gamma}_L = [m(n-m)h_n]^{-1} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \{I(Y_j - Y_i > \hat{\beta}_L) - I(Y_j - Y_i > \hat{\beta}_U)\} \tag{3.12}$$

$$= [m(n-m)h_n]^{-1} \sum_{i=1}^{m} \sum_{j=m+1}^{n} I(|Y_j - \hat{\beta} - Y_i| < h_n/2) .$$

Thus $\hat{\gamma}_L$ is like a window estimate computed on the residuals after fitting the linear model. In fact, we can write

$$\hat{\gamma}_L = \int f_m(x) dF_{n-m}(x) \tag{3.13}$$

where $f_m(\cdot)$ is a rectangular window estimate of the density based on the first sample and $F_{n-m}(\cdot)$ is the empirical cdf based on the residuals of the second sample. Similarly,

$$\hat{\gamma}_L = \int f_{n-m}(x)dF_m(x), \tag{3.14}$$

where the density estimate is computed on residuals from the second sample and $F_m(\cdot)$ is the empirical cdf of the first sample.

Let $r_i = Y_i$, $i = 1, \ldots m$ and $r_i = Y_i - \hat{\beta}$, $i = m+1, \ldots, n$ denote the residuals. Then the density estimate of γ based on $r_1, \ldots r_n$ is $\int f_n(x)dF_n(x)$, which we can write as

$$\hat{\gamma} = [n^2h_n]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} I(|r_j - r_i| < h_n/2) \tag{3.15}$$

$$= [n^2h_n]^{-1}[2 \sum_{i=1}^{m} \sum_{j=m+1}^{n} I\{|Y_j - \hat{\beta} - Y_i| < h_n/2\}$$

$$+ \sum_{i=1}^{m} \sum_{j=1}^{m} I\{|Y_j - Y_i| < h_n/2\}$$

$$+ \sum_{i=m+1}^{n} \sum_{j=m+1}^{n} I\{|Y_j - Y_i| < h_n/2\}].$$

Hence, from 3.12,

$$\hat{\gamma} = \frac{2\ m(n-m)}{n^2} \hat{\gamma}_L + \frac{m^2}{n^2} \int f_m(x)dF_m(x) \tag{3.16}$$

$$+ \frac{(n-m)^2}{n^2} \int f_{n-m}(x)d\ F_{n-m}(x)\ .$$

The density estimate $\hat{\gamma}$, constructed from the residuals, is a weighted sum of three estimates: $\hat{\gamma}_L$ (the confidence interval estimate), and two separate density estimates of $\int f^2(x)dx$ based on the two samples. It would appear that $\hat{\gamma}$ uses more information from the data than $\hat{\gamma}_L$. In the final section we discuss briefly the extension to the general linear model.

4. Extensions to the Linear Model

We again consider the comments in the last paragraph of the Introduction. In order to construct tests based on 1.9 and 1.10, without making the assumption of symmetry of $f(\cdot)$, we propose to use the ideas suggested in Section 3. In particular, for Wilcoxon scores, we would use $\hat{\gamma}$ defined in 3.15, where $r_i = Y_i - x_i'\hat{\beta}$ $i=1, \ldots, n$ and $\hat{\beta}$ is the full-model rank-estimate proposed by Jaeckel and Jureckova.

Since $r_1, \ldots, r_n$ are neither independent nor identically distributed the consistency of $\hat{\gamma}$ in 3.15 does not follow from the results cited in the previous sections. In the case of Wilcoxon scores, Aubuchon (1982) proved consistency of $\hat{\gamma}$ and studied its behavior. Further discussion, with an application to a data set, can be found in Aubuchon and Hettmansperger (1982). In 1983 a rank-regression command will be available in the Minitab statistical computing system. The output will contain the rank estimates of Jaeckel and Jureckova, and the tests 1.9 and 1.10 using the density estimate.

In a designed experiment or a regression model with replicates, the estimate $\hat{\gamma}_L$, 3.11, based on the confidence interval of a shift parameter between groups of replicates, does not require symmetry of $f(\cdot)$. A final estimate of γ is constructed by pooling the individual estimates formed from pairs of replicate groups. Draper (1981) studied these estimates in detail. However, $\hat{\gamma}$ in 3.16 suggests that a density estimate formed from all the residuals may be more informative. No careful comparison of the two approaches has been carried out yet.

References

Adichie, J. N. (1978), "Rank tests of sub-hypotheses in the general linear regression," Ann. Stat. 5, 1012-1026.

Ahmad, I. A. (1976), "On asymptotic properties of an estimate of a functional of a probability density," Scand. Actuarial J. 4, 178-181.

Antille, A. (1972), "Linearité asymptotique d'une statistique de rang," Z. Wahrschienlichkeitstheorie verw. 24, 309-324.

Antille, A. (1974), "A linearized version of the Hodges-Lehmann estimator," Ann. Stat. 2, 1308-1311.

Aubuchon, J. C. (1982), "Rank Tests in the Linear Model: Asymmetric Errors," Unpublished Ph.D. Thesis, Department of Statistics, The Pennsylvania State University.

Aubuchon, J. C. and Hettmansperger, T. P. (1982), "On the use of rank tests and estimates in the linear model," Technical Report #41, Department of Statistics, The Pennsylvania State University.

Bean, S. J. and Tsokos, C. P. (1980), "Developments in nonparametric density estimation," Int. Statist. Review 43, 267-287.

Bean, S. J. and Tsokos, C. P. (1982), "Bandwidth selection procedures for kernel density estimates," Comm. Statist.-Theor. Meth., 11(9), 1045-1069.

Bhattacharyya, B. K. and Roussas, G. G. (1969), "Estimation of a certain functional of a probability density," Skandinavisk Akturarietidskrift, 201-206.

Cheng, K. F. and Serfling, R. J. (1981), "On estimation of a class of efficacy-related parameters," Scand. Actuarial J., 83-92.

Draper, D. (1981), "Rank based robust analysis of linear models," unpublished Ph.D. Thesis, Department of Statistics, University of California, Berkeley.

Hajek, J. and Sidak, Z. (1967), Theory of Rank Tests, Academic Press, New York.

Hampel, F. R. (1974), "The influence curve and its role in robust estimation," J. Amer. Statist. Assoc., 69, 383-393.

Hodges, J. L., Jr. and Lehmann, E. L. (1963), "Estimation of location based on rank tests," Ann. Math. Statist., 34, 598-611.

Huber, P. J. (1970), "Studentizing robust estimates," Nonparametric Techniques in Statistical Inference, Ed. M. L. Puri, 453-464.

Huber, P. J. (1981), *Robust Statistics*, John Wiley, New York, New York.

Jaeckel, L. A. (1971), "Robust estimates of location: Symmetry and asymmetric contamination," *Ann. Math. Statist.*, *42*, 1020-1034.

Jaeckel, L. A. (1972), "Estimating regression coefficients by minimizing the dispersion of the residuals," *Ann. Math. Statist.*, *43*, 1449-1458.

Jureckova, J. (1971), "Nonparametric estimate of regression coefficients," *Ann. Math. Statist.*, *42*, 1328-38.

Lehmann, E. L. (1963), "Nonparametric confidence intervals for a shift parameter," *Ann. Math. Statist*. *34*, 1507-1512.

Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, Holden-day, San Francisco.

McKean, J. W. and Hettmansperger, T. P. (1976), "Tests of hypotheses based on ranks in the general linear model," *Commun. Statist.-Theor. Meth.*, *A5(8)*, 693-709.

McKean, J. W. and Hettmansperger, T. P. (1978), "A robust analysis of the general linear model based on one step R-estimates," *Biometrika* *65*, 571-579.

Parzen, E. (1962), "On estimation of a probability density function and mode," *Ann. Math. Statist*. *33*, 1065-1076.

Rosenblatt, M. (1956), "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist*. *27*, 832-837.

Ryan, T.A., Jr., Joiner, B. L. and Ryan, B. F. (1981), *Minitab Reference Manual*, Minitab Project, Pennsylvania State University, University Park, PA.

Schuster, E. (1974), "On the rate of convergence of an estimate of a functional of a probability density," *Scand. Acturial J.*, *1*, 103-107.

Schweder, T. (1975), "Window estimation of the asymptotic variance of rank estimators of location," *Scand. J. of Statist*., *2*, 113-126.

Sen, P. K. (1966), "On a distribution-free method of estimating asymptotic efficiency of a class of nonparametric tests," *Ann. Math. Statist.*, *37*, 1759-1770.

Sen, P. K. and Puri, M. L. (1977), "Asymptotically Distribution-Free Aligned Rank Order Tests for Composite Hypotheses for General Multivariate Linear Models, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, *39*, 175-186.

Sievers, G. L. (1982), "A consistent estimate of a nonparametric scale parameter," Mimeo Series #1501, Institute of Statistics, University of North Carolina.

van Eeden, C. (1972), "An analogue, for signed rank statistics of Jureckova's asymptotic linearity theorem for rank statistics," *Ann. Math. Statist*., *43*, 791-802.

Wegman, E. J. (1972), "Nonparametric probability density estimation: I. A summary of available methods," Technometrics 14, 533-546.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Pennsylvania State University Technical Report #42	2. GOVT ACCESSION NO. AD-A116432	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Rank-Based Inference Without Symmetric Errors		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) James C. Aubuchon, Pennsylvania State University Thomas P. Hettmansperger, Pennsylvania State University		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0741
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics The Pennsylvania State University University Park, PA 16802		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR042-446
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics and Probability Program Code 436 Arlington, VA 22217		12. REPORT DATE June 1982
		13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

To be published in the Proceedings of Miniconference in Nonparametric Statistics, Lexington, KY, 1982.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

rank tests, linear model, density estimates

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Statistical inference based on ranks is reviewed. The role of the parameter $\gamma = \int f^2(x)dx$ and methods for its estimation are discussed. In particular, the use of density estimation methods is shown to provide a consistent estimate γ of γ without the assumption of symmetry of the underlying distribution. The use of γ in constructing hypothesis tests in the linear model without assuming symmetry is discussed.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

D

FIL

07